# Identifying Missing Hierarchical Relations in SNOMED CT from Logical Definitions Based on the Lexical Features of Concept Names

Olivier Bodenreider
U.S. National Library of Medicine
National Institutes of Health
Bethesda, Maryland, USA
olivier.bodenreider@nih.gov

***Abstract*—Objectives. To identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. Methods. We first create logical definitions from the lexical features of concept names, which we represent in OWL EL. We infer hierarchical (*subClassOf*) relations among these concepts using the ELK reasoner. Finally, we compare the hierarchy obtained from lexical features to the original SNOMED CT hierarchy. We review the differences manually for evaluation purposes. Results. Applied to 15,833 disorder and procedure concepts, our approach identified 559 potentially missing hierarchical relations, of which 78% were deemed valid. Conclusions. This lexical approach to quality assurance is easy to implement, efficient and scalable.**

***Keywords—description logics; SNOMED CT; quality assurance; lexical features.***

## I. Introduction

Quality assurance of large biomedical terminologies remains an active area of research [1]. For example, recent investigations of SNOMED CT have highlighted issues in its hierarchical structure and demonstrated their detrimental consequences (e.g., [2]).

Both lexical features and logical definitions have been used for quality assurance purposes. Approaches based on **lexical features** generally exploit the presence of specific words in SNOMED CT terms or contrast sets of words for terms across concepts to suggest relations among concepts (e.g., [3-6]). For example, the concepts *Asthma* and *Acute asthma* can be represented by the sets of words {asthma} and {acute, asthma}, respectively. Since {asthma} is a proper subset of {acute, asthma}, the principles of lexical semantics suggest that *Acute asthma* is more specific than *Asthma* [7]. Approaches based on **logical definitions** often rely on a description logics reasoner for analyzing the facts in the ontology (e.g., [8]). The logical definitions found in SNOMED CT are sets of axioms (facts), i.e., logical statements relating concepts through "roles" (relationships), representing biomedical knowledge. For example, the axiom "Acute asthma, Clinical Course, Sudden onset AND/OR short duration" is part of the logical definition of *Acute asthma* and provides a formal representation of the acute aspect of the disease. Although logical definitions generally rely on knowledge associated with concepts, we exploit the fact that such definitions can also be created from lexical features.

The objective of this investigation is to identify missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. More specifically, we propose to leverage description logics for representing the lexical features of concept names and infer hierarchical relations based on these lexical features with a reasoner. The hierarchical relations inferred from lexical features but not present in SNOMED CT are candidates for missing relations.

## II. Background

### A. SNOMED CT

Developed by the International Health Terminology Standard Development Organization (IHTSDO), SNOMED CT is the world's largest clinical terminology. With 320,000 active concepts, it provides broad coverage of clinical medicine, including findings, diseases, and procedures for use in electronic medical records [9].

SNOMED CT provides a preferred name and synonyms for each concept ("descriptions" in SNOMED CT parlance). The "fully specified name" is guaranteed to be unique for each concept and consists of the preferred term followed by a semantic tag (e.g., *Blepharorrhaphy (procedure)* (388008)). In addition to names, all concept have a logical definition, based on definitional characteristics of the concept (not on the lexical features of the concept names). For example,

Class: *Blepharorrhaphy*
EquivalentTo:
*Suture of eyelid*
and (***Method*** some *Closure - action*)
and (***Procedure site - Direct*** some *Structure of palpebral fissure*)
and (***Using device*** some *Surgical suture, device*)

In SNOMED CT, the logical definitions are processed with a description logic reasoner for consistency validation and to

generate the hierarchical structure by inferring ***subClassOf*** relations among the concepts.

The version of SNOMED CT used in this work is the U.S. edition dated March 2016.

### B. Description logics

Description logics (DL) are a family of knowledge representation languages often used as ontology languages, and defined as a trade-off between expressivity and tractability [10]. Reasoners are computer programs that can check the consistency of the facts asserted in the ontology and infer relations among ontology classes based on these facts (i.e., infer hierarchical (***subClassOf***) relations).

Among the various flavors of DL languages available, the EL family offers sufficient expressivity for the simple definitions resulting from lexical features, as well as scalability to a large number of classes [11]. The reasoners developed for EL (e.g., ELK [12]) offer impressive performance.

As illustrated above, SNOMED CT relies on DL for representing the logical definitions it provides for its concepts. It also makes use of a reasoner for testing the consistency of these definitions across the whole ontology, as well as for inferring the hierarchy of concepts. In this work, we apply the reasoner not to the logical definitions provided by SNOMED CT to represent biomedical knowledge, but rather to the definitions we generate from the lexical features of the terms of SNOMED CT concepts.

### C. Quality assurance of biomedical ontologies

Approaches to quality assurance in biomedical ontologies can be classified into lexical, structural and semantic approaches [13]. Lexical approaches rely on the lexical features of terms; structural approaches analyze the hierarchical structure of ontologies; and semantic approaches exploit the relations among concepts (including logical definitions). Examples of lexical and semantic approaches applied to quality assurance in SNOMED CT were presented earlier in the introduction. (Structural approaches are less relevant to this work and will not be discussed here.)

Of note, while DL techniques are generally used in the context of semantic approaches, in this work, we leverage a DL reasoner for the implementation of a lexical approach to QA, since our logical definitions are created on the basis of lexical features.

The compositionality of terms in biomedical ontologies is well documented and has been exploited for quality assurance purposes (e.g., [14, 15]). However, Mungall used *ad hoc* programming (in Prolog) rather than a DL reasoner to infer relations among terms. Our approach is also much simpler in that it only relies on sets of words and only attempts to elicit hierarchical relations.

### D. Specific contribution

The specific contribution of this work is not in leveraging the compositionality of biomedical terms for suggesting relations, but rather in proposing a description logics approach to doing so. While *ad hoc* programming is usually necessary for comparing bags of words, our work demonstrates it can also be supported effectively by a DL reasoner. To our knowledge, this is the first attempt to generate logical definitions based on the lexical features of concept names in SNOMED CT for quality assurance purposes.

## III. Methods

Our method for identifying missing hierarchical relations from SNOMED CT can be summarized as follows. We first create logical definitions from the lexical features of concept names, which we represent in the web ontology language, OWL. We infer hierarchical (***subClassOf***) relations among these concepts using a reasoner. Finally, we compare the hierarchy obtained from lexical features to the original SNOMED CT hierarchy. We review the differences manually for evaluation purposes. In this preliminary investigation, we applied this approach to a significant subset of the *Clinical Finding* hierarchy rooted with the concept *Disorder of head (disorder)* (118934005) and a smaller subset of the *Procedure* hierarchy rooted with the concept *Operative procedure on head (procedure)* (89901005).

### A. Creating logical definitions based on the lexical features of concept names

For each concept under investigation, we extract the fully specified name, which consists of the preferred term (e.g., "*Disorder of head*") followed by a semantic tag in parentheses (e.g. "*disorder*"). For each concept *C* with fully specified name "$w_1\ w_2 \ldots w_n\ (T)$", where $\{w_1, w_2, \ldots w_n\}$ is the set of words in the preferred term and where *T* is the semantic tag, we create a logical definition of the following form (expressed in the simplified OWL syntax known as Manchester syntax [16]):

Class: *C*
  EquivalentTo:
    *T*
    and (***has_word*** some $w_1$)
    and (***has_word*** some $w_2$)
    …
    and (***has_word*** some $w_n$)

For example, the class definition for the concept *Complete ablepharon (disorder)* (708541009) is shown in Fig. 1.

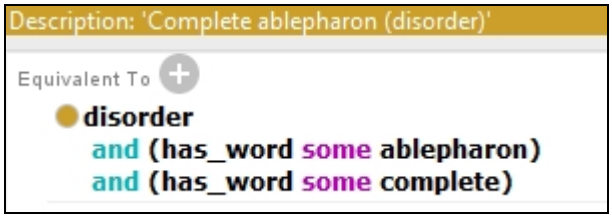


Fig. 1. Class definition for the concept *Complete ablepharon (disorder)*

In practice, we use a simple script to create an OWL file that contains the class definitions for all the concepts under investigation. The words "the" and "of", present in a large

proportion of terms, are omitted when generating the class definitions.

Of note, the OWL constructs used in these definitions (namely class equivalence and existential quantification to a class expression) are compatible with the OWL 2 EL profile [11].

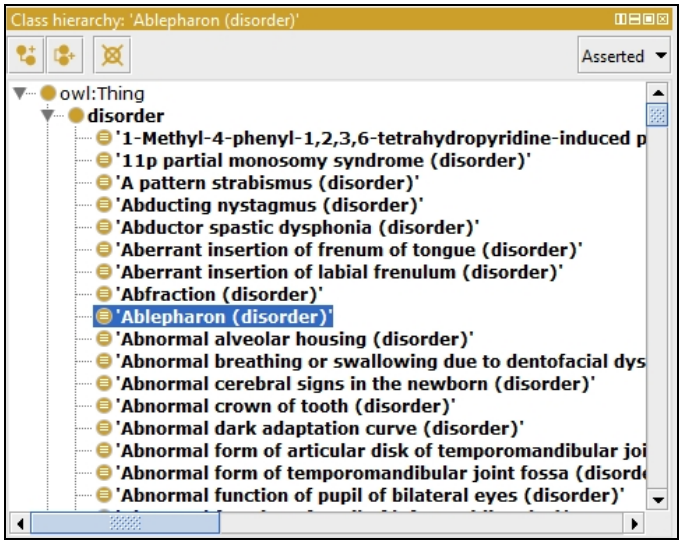


Fig. 2. Asserted hierarchy – *Ablepharon (disorder)* prior to running the reasoner (no inferred subclasses)

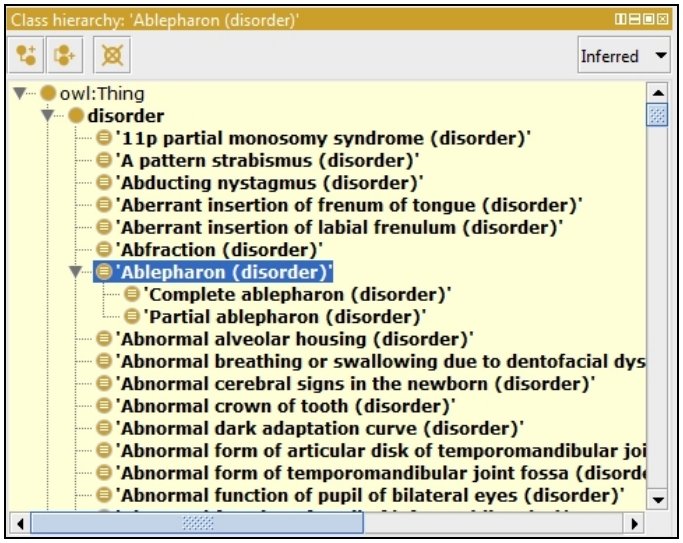


Fig. 3. Inferred hierarchy – *Ablepharon (disorder)* after the reasoner has run (two inferred subclasses: *Complete ablepharon (disorder)* and *Partial ablepharon (disorder)*)

### *B. Inferring subClassOf relations from lexical features*

We load this OWL file in the Protégé ontology editor (5.0 beta), in which we have installed the plugin for the ELK reasoner [12], specially optimized for classifying OWL 2 EL ontologies. Prior to running the reasoner, the SNOMED CT concepts imported into Protégé appear as a flat list (i.e., with no hierarchical structure) under the classes created for the semantic tags (Fig. 2). After ELK has run, inferred ***subClassOf*** axioms among the SNOMED CT concepts have been added to the ontology and the concepts are no longer displayed as a flat list (Fig. 3). For example, the three concepts *Ablepharon (disorder)* (13401001), *Complete ablepharon (disorder)* (708541009), and *Partial ablepharon (disorder)* (45484000) are listed under *disorder* in the asserted hierarchy (Fig. 2), but *Complete ablepharon (disorder)* and *Partial ablepharon (disorder)* are subclasses of *Ablepharon (disorder)* in the inferred hierarchy (Fig. 3).

Since the ***subClassOf*** relations are inferred from lexical features, we need to filter out complex terms with prepositional phrases to avoid generating wrong ***subClassOf*** relations. For example, for *Dementia due to Parkinson's disease (disorder)* (101421000119107), a ***subClassOf*** relation is inferred to both *Dementia (disorder)* (52448006) and *Parkinson's disease (disorder)* (49049000). Similarly, for *Goniopuncture without goniotomy (procedure)* (202727004), a ***subClassOf*** relation is inferred to both *Goniopuncture (procedure)* (265293008) and *Goniotomy (procedure)* (265292003). While this behavior is expected from the reasoner, it is not desirable, because *Dementia due to Parkinson's disease (disorder)* is not a kind of *Parkinson's disease (disorder)* as suggested by the prepositional expression "due to". Similarly, *Goniopuncture without goniotomy (procedure)* specifically excludes *Goniotomy (procedure)*. In practice, to avoid generating such wrong ***subClassOf*** relations, we filter out the relations generated when the name of the most specific ("child") concept contains any of the following words: "and", "or", "and/or", "with", "without", "from", "due to", "secondary to", "except", "by", "after", "revision" and "ligation for".

### *C. Comparing the hierarchy inferred from lexical features to the original hierarchy*

To analyze which relations from the inferred hierarchy are not already in the original SNOMED CT hierarchy (i.e., the hierarchy found in the SNOMED CT distribution), we need to generate these two sets of hierarchical relations and compute the difference between them. Using Protégé, we export the inferred ***subClassOf*** axioms to a file in RDF format for comparison to the original hierarchical relations in SNOMED CT. Using a simple script, we write the original hierarchical relations in SNOMED CT to RDF for the subhierarchies under investigation. In practice, because the inferred relations can be between any two classes, we enrich the original hierarchy with the transitive closure of ***subClassOf*** relations. We load the files for the two sets of relations, inferred and original, into the triple store Virtuoso and use a SPARQL query to compute the set of hierarchical relations from the inferred set that is not part of the hierarchical relations originally in SNOMED CT (transitively closed). The SPARQL 1.1 operator MINUS makes such comparison between two graphs extremely easy.

### *D. Evaluation*

We manually review for validity a random subset of 100 inferred relations that are not present in the original SNOMED CT hierarchy (transitively closed).

## IV. Results

### A. Creating logical definitions based on the lexical features of concept names

We created logical definitions based on the lexical features of concept names for the 12,088 concepts (4871 distinct words) of the subhierarchy rooted with the concept *Disorder of head (disorder)* (118934005) and for the 3795 concepts (1899 distinct words) of the subhierarchy rooted with the concept *Operative procedure on head (procedure)* (89901005).

### B. Inferring subClassOf relations from lexical features

Running the ELK reasoner took a few seconds and resulted in the creation of 7079 inferred ***subClassOf*** relations among the concepts of the subhierarchy rooted with the concept *Disorder of head (disorder)*. Similarly, 1357 relations were inferred in the subhierarchy rooted with the concept *Operative procedure on head (procedure)*.

### C. Comparing the hierarchy inferred from lexical features to the original hierarchy

After subtracting from the inferred ***subClassOf*** relations created by the reasoner those ***subClassOf*** relations already present in the original version of SNOMED CT (transitively closed), we obtained 1210 inferred ***subClassOf*** relations for the *Disorder of head (disorder)* hierarchy and 242 inferred ***subClassOf*** relations for the *Operative procedure on head (procedure)* hierarchy. Of these, 469 ***subClassOf*** relations for disorders and 90 for procedures met our criteria for review (i.e., the name of the child concept does not contain any of the prepositional and other expressions listed earlier).

### D. Evaluation

The random subset of 100 inferred ***subClassOf*** relations we reviewed comprises 83 disorders and 17 procedures. Overall, 78 relations were deemed valid, 19 invalid and 3 questionable (i.e., these relations seem to have face validity, but may not be compliant with SNOMED CT editorial policies). Examples of such relations are listed in Table I.

## V. Discussion

### A. Findings

As expected, a vast majority of the hierarchical relations suggested lexically were already present in the original SNOMED CT hierarchy (transitively closed). Specifically, only 1210 of the 7079 hierarchical relations for disorders (17%) and 242 of the 1357 hierarchical relations for procedures (18%) were not already represented in SNOMED CT.

However, it was somewhat surprising to us to see that a large number of potentially missing hierarchical relations had been generated from this simple technique based on lexical features. Assuming 80% of the 559 hierarchical relations generated are correct, we discovered 447 missing hierarchical relations among the 15,883 concepts under investigation. Interestingly, the proportion is roughly the same for disorders and procedures.

In addition to the evaluation, we performed a cursory review of the 559 potentially missing hierarchical relations, among which we identified a few patterns. In 31 cases, the missing relation was between "carcinoma in situ of <some anatomical structure>" and "carcinoma of <some anatomical structure>" (or "<some anatomical structure> carcinoma"), for example, between *Carcinoma in situ of palate (disorder)* (92670007) and *Palate carcinoma (disorder)* (274084007). Another such patterns was found in 23 cases between "congenital <some disorder>" and the unqualified disorder, for example, between *Congenital anterior staphyloma (disorder)* (253230008) and *Anterior staphyloma (disorder)* (231888000).

### B. Technical significance

The novel aspect of this work is to use a DL approach to lexical similarity. In practice, it means that no *ad hoc* programming is required for identifying partial ordering relations among sets of words for terms in an ontology reflecting hierarchical relations among the corresponding concepts. Instead, logical definitions created from lexical features can simply be represented in DL formalism and run through a reasoner to infer the relevant ***subClassOf*** relations. As shown here, this approach is easy to implement, efficient and scalable. The only programming required is for serializing the logical definitions in the appropriate DL format.

Moreover, given that SNOMED CT already uses DL techniques for representing its logical definitions based on biomedical knowledge and an EL reasoner for inferring its hierarchy, it can be expected that the IHTSDO could easily integrate the lexical approach to quality assurance proposed here.

Finally, having two kinds of logical definitions (from biomedical knowledge and from lexical features) represented in the same formalism would make it possible to integrate them into the same framework, for example to test the consistency between the two kinds of definitions.

### C. Limitations and future work

This preliminary investigation is limited to two subhierarchies of SNOMED CT for diseases and procedures. However, we also generated definitions and inferred hierarchy for the whole SNOMED CT and did not notice any scalability issues. We did not leverage SNOMED CT synonyms for creating logical definitions, but this should be a natural extension of this investigation. In future work, we also would like to normalize terms before creating the definitions, since normalization is common approach to managing term variation [17].

This bag-of-word approach to comparing terms tends to generate more false positives than a linguistically motivated approach, where the head of the noun phrase would be required to be the same in two hierarchically related concepts, as we did in other work [18]. In fact, many of the errors detected during the evaluation correspond to cases where the specific term is linked to a term that does not contain the head of the noun phrase of the specific term. However, the bag-of-word approach is much easier to implement than linguistically

motivated approaches, and we showed that false positives can be mitigated in part by filtering out complex terms.

In this preliminary investigation, we performed a limited evaluation. Given the encouraging results, we plan to extend the investigation to the entirety of SNOMED CT, evaluate the results more thoroughly, and share them with the SNOMED CT developers at the IHTSDO.

Finally, the lexical approach to quality assurance proposed here could also complement structural approaches, such as the lattice-based approach we proposed earlier [19].

### *D. Generalization*

This approach to identifying missing hierarchical relations would be applicable not only to the entirety of SNOMED CT, but to other biomedical ontologies as well. More specifically, it could be applied to any biomedical ontology for which concept names and hierarchical relations are available (i.e., most ontologies). The same approach could also be applied to the creation of partial mappings.

## ACKNOWLEDGMENT

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. We would like to thank Dr. GQ Zhang for providing motivation and encouragement for this investigation.

## REFERENCES

[1] J. Geller, et al., "Special issue on auditing of terminologies," J Biomed Inform, vol. 42, no. 3, 2009, pp. 407-411.

[2] A.L. Rector, et al., "Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications," J Am Med Inform Assoc, vol. 18, no. 4, 2011, pp. 432-440.

[3] O. Bodenreider, et al., "Assessing the consistency of a biomedical terminology through lexical knowledge," Int J Med Inform, vol. 67, no. 1-3, 2002, pp. 85-95.

[4] K.E. Campbell, et al., "A "lexically-suggested logical closure" metric for medical terminology maturity," Proc AMIA Symp, 1998, pp. 785-789.

[5] E. Mikroyannidi, et al., "Analysing Syntactic Regularities and Irregularities in SNOMED-CT," J Biomed Semantics, vol. 3, no. 1, 2012, pp. 8.

[6] E. Pacheco, et al., "Detecting Underspecification in SNOMED CT concept definitions through natural language processing," AMIA Annu Symp Proc, vol. 2009, 2009, pp. 492-496.

[7] D.A. Cruse, *Lexical semantics*, Cambridge University Press, 1986, p. xiv, 310.

[8] K. Dentler and R. Cornet, "Intra-axiom redundancies in SNOMED CT," Artif Intell Med, vol. 65, no. 1, 2015, pp. 29-34.

[9] IHTSDO, "SNOMED CT," 2016.

[10] F. Baader, et al., "Description logics," *Handbook on ontologies*, International handbooks on information systems, S. Staab and R. Studer, eds., Springer, 2004, pp. 3-28.

[11] W3C, "OWL 2 Web Ontology Language Profiles (Second Edition)," 2012; https://www.w3.org/TR/owl2-profiles/#OWL_2_EL.

[12] Y. Kazakov, et al., "The Incredible ELK: From Polynomial Procedures to Efficient Reasoning with EL Ontologies," Journal of Automated Reasoning, vol. 53, no. 1, 2013, pp. 1-61.

[13] X. Zhu, et al., "A review of auditing methods applied to the content of controlled biomedical terminologies," J Biomed Inform, vol. 42, no. 3, 2009, pp. 413-425.

[14] C.J. Mungall, "Obol: integrating language and meaning in bio-ontologies," Comp Funct Genomics, vol. 5, no. 6-7, 2004, pp. 509-520.

[15] P.V. Ogren, et al., "The compositional structure of Gene Ontology terms," Pac Symp Biocomput, 2004, pp. 214-225.

[16] W3C, "https://www.w3.org/TR/owl2-manchester-syntax/," 2012.

[17] A.T. McCray, et al., "Lexical methods for managing variation in biomedical terminologies," Proc Annu Symp Comput Appl Med Care, 1994, pp. 235-239.

[18] F. Dhombres and O. Bodenreider, "Interoperability between phenotypes in research and healthcare terminologies--Investigating partial mappings between HPO and SNOMED CT," J Biomed Semantics, vol. 7, 2016, pp. 3.

[19] G.Q. Zhang and O. Bodenreider, "Large-scale, Exhaustive Lattice-based Structural Auditing of SNOMED CT," AMIA Annu Symp Proc, vol. 2010, 2010, pp. 922-926.

TABLE I. EXAMPLES OF SUBCLASSOF RELATIONS INFERRED FROM LEXICAL FEATURES

| Hierarchy | Child ID | Child name | Parent ID | Parent name | Valid |
|---|---|---|---|---|---|
| Procedure | 239405007 | Alveolar bone graft ***to mandible*** (procedure) | 178493006 | Alveolar bone graft (procedure) | yes |
| Disorder | 402819001 | ***Basal cell*** carcinoma ***of skin*** of lip (disorder) | 269515006 | Carcinoma of lip (disorder) | yes |
| Disorder | 92670007 | Carcinoma ***in situ of*** palate (disorder) | 274084007 | Palate carcinoma (disorder) | yes |
| Disorder | 232225005 | Chronic ***bacterial*** otitis externa (disorder) | 53295002 | Chronic otitis externa (disorder) | yes |
| Disorder | 700278007 | ***Congenital*** vascular anomaly of eyelid (disorder) | 69973000 | Vascular anomaly of eyelid (disorder) | yes |
| Procedure | 31230008 | ***Electrocoagulation*** of retina ***for*** repair ***of tear*** (procedure) | 450698009 | Repair of retina (procedure) | yes |
| Disorder | 40571009 | Hallucinogen intoxication ***delirium*** (disorder) | 50320000 | Hallucinogen intoxication (disorder) | no |
| Disorder | 609209009 | ***Infection of*** preauricular sinus (disorder) | 204271000 | Preauricular sinus (disorder) | no |
| Disorder | 237664006 | Pituitary stalk compression ***hyperprolactinemia*** (disorder) | 237723009 | Pituitary stalk compression (disorder) | no |
| Procedure | 440303005 | Suture of ***tongue to*** lip ***for micrognathia*** (procedure) | 3889008 | Suture of lip (procedure) | no |